Reading and Writing (2021) 34:2461–2480
https://doi.org/10.1007/s11145-021-10153-6

Validity of automated text evaluation tools for written-expression curriculum-based measurement: a comparison study

Milena A. Keller-Margulis[1] · Sterett H. Mercer[2] · Michael Matta[1]

Accepted: 21 March 2021 / Published online: 1 April 2021

Abstract

Existing approaches to measuring writing performance are insufficient in terms of both technical adequacy as well as feasibility for use as a screening measure. This study examined the validity and diagnostic accuracy of several approaches to automated text evaluation as well as written expression curriculum-based measurement (WE-CBM) to determine whether an automated approach improves technical adequacy. A sample of 140 fourth grade students generated writing samples that were then scored using traditional and automated approaches and examined in relation to the statewide measure of writing performance. Results indicated that the validity and diagnostic accuracy for the best performing WE-CBM metric, correct minus incorrect word sequences, and the automated approaches to scoring were comparable, with automated approaches offering potentially improved feasibility for use in screening. Averaging scores across three time points was necessary, however, in order to achieve improved validity and adequate levels of diagnostic accuracy across the scoring approaches. Limitations, implications, and directions for future research regarding the use of automated scoring approaches for screening are discussed.

Keywords Written expression · Curriculum-based measurement · Automated text evaluation · Validity

Introduction

Advanced writing skills are critical for educational and occupational success (Perin, 2020; Roebuck et al., 1995; Stevens, 2005). Unfortunately, there is increasing evidence to suggest that many students do not demonstrate adequate levels

✉ Milena A. Keller-Margulis
mkmargulis@central.uh.edu

1 University of Houston, Houston, USA

2 University of British Columbia, Vancouver, Canada

of performance in the area of writing (National Center for Educational Statistics, 2012). Efforts to improve writing skills hinge on the ability to identify those students who are at-risk for poor performance and to measure their response to changes in instruction (McMaster et al., 2020); however, existing approaches to scoring student writing samples are either too elaborate for use in screening or do not offer the technical adequacy required for decision-making about student risk status (e.g., Gansle et al., 2002; McMaster & Espin, 2007). Advances in automated text evaluation may offer a viable alternative to address these measurement limitations (Mercer et al., 2019). The purpose of this study was to examine and compare the validity and diagnostic accuracy of several automated approaches to scoring written expression screening samples to determine whether automated scoring yields scores with improved technical adequacy.

Measuring writing

There are numerous approaches to measuring writing, but few that are conducive to regular screening and progress monitoring. Curriculum-based measurement (CBM) is an approach characterized by its reliability and validity for making decisions about student performance, as well as its brevity and ease of implementation by various users (Deno, 1985). Written expression CBM (WE-CBM) requires students to generate written material in response to a prompt or a picture; in the early grades (e.g., kindergarten and first grade), WE-CBM includes producing words or simple sentences (Ritchey et al., 2016), whereas, in the upper elementary grades, WE-CBM often includes writing a story with 1 min for the student to plan and 3 or 5 min to complete the writing task (Hosp et al., 2016). Student writing samples are then scored for various indicators of performance.

WE-CBM scoring metrics include total words written (TWW), words spelled correctly (WSC), and correct word sequences (CWS), which provide the total number of words in the sample, an indication of spelling accuracy, and a sentence-level indicator of grammar and syntax, respectively. These metrics are considered production-dependent as they are derived from the amount of text a student produces during the time allowed. Although they may be informative, it is difficult to use these metrics to compare writing quality when students produce samples of varying length. The need to further extended scoring metrics resulted in the creation of the accurate production index of correct minus incorrect word sequences (CIWS; Espin et al., 2000) as a way to capture both accuracy and fluency of writing.

Early research on WE-CBM metrics indicated sufficient technical adequacy for the simpler production-dependent metrics (TWW and CWS; McMaster & Espin, 2007); however, recent meta-analytic findings demonstrated greater validity in relation to standardized or statewide writing achievement tests for more complex metrics (CWS and CIWS; Romig et al., 2017) that require more time to score (Gansle et al., 2002). Specifically, the mean-weighted criterion validity coefficients were $r=0.51$ for CWS and 0.61 for CIWS, compared to $r=0.37$ for TWW and 0.44 for WSC (Romig et al., 2017). In fact, CIWS outperformed the other metrics across grade levels. The procedures used in the meta-analysis involved selection of the highest

coefficient reported in the individual studies. As a result, the findings reported represent the highest possible coefficients and the inclusion of a broader representation of coefficients could ultimately lower results.

Despite this evidence for the validity of WE-CBM scores, recent research has questioned the adequacy of WE-CBM scores as estimates of student writing skill when based on typical administration procedures. Although there is evidence of adequate alternate-form reliability across WE-CBM prompts (Keller-Margulis et al., 2016; McMaster & Espin, 2007) and adequate interscorer reliability (McMaster & Espin, 2007), the use of a single 3-min writing sample per student may not be sufficient for reliably estimating student writing skill. In a generalizability theory study of second-fifth grade students, Keller-Margulis et al. (2016) found that WE-CBM scores based on a single 3-min writing sample per student would yield reliability coefficients of 0.50–0.63 for relative decisions; adequate reliability (≥ 0.80) could be achieved when WE-CBM scores were based on three 3-min samples or two longer duration samples of 4–7 min, depending on grade level. Also demonstrating the need for multiple writing samples per student, Kim et al. (2017), in a generalizability theory study of students in third and fourth grade, found that two to four 15-min samples were needed for adequate reliability for relative decisions. Collectively, these reliability and validity studies indicate that more complex scoring of multiple, longer duration samples will be needed to use WE-CBM for universal screening decisions; however, these requirements raise concerns about the feasibility of using WE-CBM for screening.

Feasibility of WE-CBM scoring

The increased scoring time required compared to simpler metrics such as TWW makes the more complex metrics of CWS or CIWS a challenging choice for screening (Payan et al., 2019). The time investment in scoring writing samples from entire grade levels or schools, for minimally acceptable validity, would likely not be considered a reasonable use of resources. Several studies indicate the potential feasibility challenges associated with scoring metrics such as CWS by examining the time required to score samples (Gansle et al., 2002; Malecki & Jewell, 2003). The estimates for CWS scoring time range from just under a minute for third and fourth graders (i.e., 57.3 s; Gansle et al., 2002) to just over a minute for third through fifth graders (i.e., 74.3 s; Malecki & Jewell, 2003). Using the estimate from Gansle et al. (2002) combined with the knowledge that multiple samples are required for reliability (Keller-Margulis et al., 2016; Kim et al., 2017), scoring three 3-min samples would require 2.9 min for one student. If there are approximately 25 students in a classroom, the total class scoring time would be about 72 min for the CWS metric. Extending the duration of writing, as suggested by findings indicating a need for longer samples to improve technical adequacy (Keller-Margulis et al., 2016; Kim et al., 2017), would likewise extend the time and resources required for scoring. For example, if 15-min samples were to be collected, it would take approximately 4.77 min to score a single sample for CWS, and nearly 120 min to score 25 samples from a classroom. In sum, the need for complex scoring of multiple, long-duration

samples presents a considerable feasibility challenge when conducting screening on an entire grade level or multiple grade levels of students.

Applying automated text evaluation to WE-CBM

Automated text evaluation may be a solution to these feasibility challenges. The idea of using computer scoring in WE-CBM is not new—for example, Espin et al. (1999) evaluated the criterion-related validity of several simple descriptive measures calculated by word processing software such as the number of characters, words, and sentences written, and Gansle et al. (2002) evaluated the validity of several readability metrics provided in word processing programs. These efforts did not yield validity coefficients above the commonly-applied threshold of $r = 0.50$ recommended for minimally sufficient validity in WE-CBM research (McMaster & Campbell, 2008) or the National Center on Intensive Intervention's (2018) more stringent standard (i.e., the lower bound of the r confidence interval ≥ 0.60); however, more sophisticated automated text evaluation options are now available.

Two studies have evaluated the validity of Project Essay Grade (PEG; Page, 2003), a commercially-available automated text evaluation tool, for universal screening in the upper elementary and middle school grades. First, Wilson et al. (2016) examined the validity of the PEG overall score and some other metrics, including TWW and WSC, calculated on 60-min persuasive writing samples in the fall with performance on a statewide writing test in the spring for sixth grade students. Results for predicting whether students met proficiency standards on the state test indicated acceptable diagnostic accuracy for a screening instrument (i.e., area under the receiver operating curve [AUC] of at least 0.75; Smolkowski et al., 2016), with AUC values at 0.78, 0.75, and 0.76 for PEG, TWW, and WSC, respectively. Second, Wilson (2018) examined the validity of PEG to score 30-min argumentative writing samples from third and fourth grade students in the fall and spring in relation to the Smarter Balanced English Language Arts (ELA) assessment, a combined assessment of reading and writing administered across multiple states. AUC values for predicting proficiency on the ELA assessment were largely similar to the findings for PEG in Wilson et al. (2016), with AUC = 0.74 and 0.75 in the fall and spring of third grade, and AUC = 0.79 and 0.83 in the fall and spring of fourth grade.

Although both studies provide evidence in support of using automated text evaluation for universal screening, there are two notable limitations to the studies. In both studies, the time limits for the screening samples (60- and 30-min) were far longer than typical practices for universal screening in WE-CBM, and in Wilson et al. (2016), the time limit for the screening samples was longer than the time allowed on the statewide writing test. Also, some aspects related to the use of PEG in the studies can be considered a limitation. Because PEG scoring is proprietary, the specific text characteristics included in scoring models and their weightings are unclear. This limited scoring model transparency has been criticized both for automated text evaluation tools (Perelman, 2014) and more broadly for socially consequential computer-based algorithmic decisions (Rainie & Anderson, 2017).

These transparency concerns are less applicable to recent efforts to develop open-source software tools that can be used for automated scoring of writing samples in universal screening; in contrast to proprietary software, the full software code is available for inspection and modification in open-source software. Consistent with recent calls for adopting open science practices in special education (Cook et al., 2018), developing and evaluating open-source tools can facilitate critical review and refinement of scoring models, replication efforts by other research teams, and greater usage of tools by removing cost barriers. For example, Mercer et al. (2019) examined the validity of scoring models based on Coh-Metrix (Graesser et al., 2014), a free, but proprietary, tool originally designed to predict reading comprehension difficulty of texts, relative to WE-CBM scores for 7-min narrative writing samples from students in second through fifth grade. Results indicated that composite scores based on Coh-Metrix could predict raters' holistic quality ratings on the screening samples, both for the samples used to generate the Coh-Metrix scores and on similar samples collected 3 months apart, and that correlations with holistic quality were similar for composites based on Coh-Metrix ($r=0.73$–0.81) and WE-CBM scores ($r=0.74$–0.77). A key limitation to this study is that performance was only evaluated in relation to holistic quality on the screening samples—there was no external criterion measure.

These Coh-Metrix based scoring models are now openly available in writeAlizer (Mercer, 2020), an R package with complete documentation of scoring models in online materials. In addition to implementing scoring models based on Coh-Metrix scores, writeAlizer also supports ReaderBench (Dascalu et al., 2018) scores as inputs; ReaderBench, similar to Coh-Metrix, is a tool designed to assess text complexity to predict reading comprehension difficulty, but ReaderBench has the additional advantage of being open source, which may further facilitate transparency and future software development. Both the Coh-Metrix and ReaderBench scoring models in writeAlizer were trained to primarily predict quality of idea development and organization on narrative writing samples; however, commercial options such as PEG typically include scores representing writing mechanics (e.g., spelling) as part of overall scores. For this reason, we investigate if adding automated indicators of spelling and grammatical errors improves criterion-related validity for the writeAlizer scoring models in the current study.

It is important to note that automated tools generate a wide range of score types, some of which are similar to WE-CBM scores and others that are quite different. For example, tools such as Coh-Metrix and ReaderBench generate text descriptive scores such as word count that are nearly identical to the TWW WE-CBM score; however, other scores are generated, such as indicators of lexical diversity, syntactic complexity, and characteristics of discourse, that have no direct analogues in WE-CBM scoring. PEG scores, which were designed to predict human-generated analytic writing rubric scores, also differ conceptually from WE-CBM scores. Despite these differences in the specific writing constructs assessed across the tools, automated scoring models all yield overall writing quality scores (directly in PEG, and through writeAlizer composite scoring models for Coh-Metrix and ReaderBench), and WE-CBM scores also were designed

to index overall writing quality. For this reason, we focus on these overall scores in our evaluation of automated scoring vs. WE-CBM for universal screening.

Current study

To advance efforts to develop free and transparent automated tools for universal screening of writing skills, the purpose of this study was to examine the criterion-related validity and diagnostic accuracy of proprietary vs. open source automated scoring of narrative samples, relative to WE-CBM hand scoring, for predicting statewide writing test performance in fourth grade students. We evaluate the performance of three automated scoring approaches, including writeAlizer, with Coh-Metrix or ReaderBench scores as inputs, and PEG. Building on prior generalizability theory studies indicating that multiple samples are needed for adequate reliability (Keller-Margulis et al., 2016; Kim et al., 2017), we also compare the validity of scores based on multiple vs. single screening samples per student. We address the following research questions:

1. Does automated scoring of narrative samples, relative to WE-CBM hand scoring, yield improved criterion-related validity and diagnostic accuracy in relation to a statewide writing test?
2. Is the performance of free automated scoring tools (writeAlizer with Coh-Metrix or ReaderBench scores) comparable to a commercial automated tool (PEG)?
3. Are criterion-related validity and diagnostic accuracy improved when automated spelling and grammar scores are added to writeAlizer?
4. Are criterion-related validity and diagnostic accuracy improved for all scoring approaches when scores are based on three vs. one writing sample per student?

Method

Participants

A total of 140 fourth-grade students (72 boys, corresponding to 51.43% of the sample) who completed the statewide writing test and at least one screening sample from two schools located in the Southwestern United States participated in the study. The majority were Hispanic (57.14%), followed by African American (34.29%), Asian (5.00%), and Caucasian (3.57%). Four students (corresponding to 2.86% of the sample) received special education services, and 53 students (37.86%) were identified as English Learners. Of the participants, 113 students (80.71%) met or exceeded the satisfactory (proficient) performance standard on the statewide writing assessment.

Measures

Traditional WE-CBM Metrics

Writing samples were hand-scored by calculating four WE-CBM metrics (Hosp et al., 2016). TWW indicates the number of total words (defined as any group of letters delimited by white spaces), regardless of spelling mistakes or their meaning. WSC is the number of words that are spelled correctly, regardless of the context. CWS is the count of two adjacent words that are spelled correctly and are acceptable within the context of the sentence. CIWS is calculated as the difference between correct and incorrect word sequences; incorrect sequences are consequent to errors in spelling, punctuation, capitalization, syntax, and semantics. Interrater reliability was calculated for all four WE-CBM metrics at each time point. TWW, WSC, and CWS showed coefficients equal to or above 0.90, and CIWS above 0.80.

Automated writing evaluation

Four approaches to automated writing evaluation were compared in this study: (a) the writeAlizer scoring model with ReaderBench scores as inputs, (b) the writeAlizer scoring model with Coh-Metrix scores as inputs, (c) writeAlizer scores combined with automated indicators of spelling and grammar, and (d) PEG total scores. Each approach is described below.

writeAlizer writeAlizer (Mercer, 2020) is an R package that generates predicted quality scores from free (Coh-Metrix and ReaderBench) and open-source tools (ReaderBench) that were originally developed to predict text complexity and reading comprehension difficulty. The writeAlizer scoring models were developed based on 7-min narrative writing samples from second- through fifth-grade students (see Mercer et al., 2019, for more details); full documentation of the scoring models are available on the writeAlizer GitHub site (Mercer, 2020).

ReaderBench ReaderBench (Dascalu et al., 2018) is an open-source tool that computes over 400 textual complexity indices that are grouped into five categories: surface features, word complexity, syntax and morphology, semantic cohesion, and discourse structure. Indices range from simple measures of written language (e.g., word and sentence length, and unique words used) to more sophisticated linguistic aspects (e.g., lexical chains and discourse connectives). To our knowledge, ReaderBench has not been evaluated as a tool to analyze elementary students' writing; however, prior research has found ReaderBench textual complexity indices calculated from undergraduates' essays to predict vocabulary and reading comprehension scores on standardized assessments (Allen et al., 2016), and the textual complexity indices have been integrated in a tool that provides automated feedback on student essays (Botarleanu et al., 2019).

Coh-Metrix Coh-Metrix (Graesser et al., 2014) is a free tool that analyzes multiple levels of written language, with an emphasis on cohesion, to predict reading comprehension difficulty. Coh-Metrix calculates 108 indices in 11 categories: descriptives, text easability principal component scores, referential cohesion, latent semantic analysis, lexical diversity, connectives, situation model, syntactic complexity, syntactic pattern density, word information, and readability. Prior work has found Coh-Metrix indices useful in predicting writing performance in elementary (Mercer et al., 2019) and middle school (Wilson et al., 2017) students.

Spelling and grammar Grammar And Mechanics Error Tool (GAMET; Crossley et al., 2019) is a free software application that counts the frequency of structural and mechanics errors in texts. We used counts of two types of errors generated by GAMET: misspellings and grammar mistakes. We divided these counts of errors by the total number of words recognized by GAMET. A study comparing errors found in GAMET on essays written by high school students with expert ratings provides some evidence that the GAMET-identified errors are accurate and meaningful (Crossley et al., 2019); however, GAMET under-identifies grammatical errors identified by expert raters. In the same study, there was a small correlation between the number of overall errors identified by GAMET and rater judgments of holistic essay quality ($r = -0.20$, $p < 0.001$).

PEG PEG (Page, 2003) is a commercially available scoring system for the assessment of writing quality for third- to twelfth-grade students. PEG uses a combination of language processing techniques to measure more than 500 variables related to writing quality (e.g., maturity of words, number of sentences per paragraph, spelling, grammar, and syntax errors, transition adverbs) that are weighted to generate scores (1 to 5 points) on six dimensions: conventions, ideas, organization, sentence structure, style, and word choice. These scores are summed to generate an overall writing quality score. Consistent with prior studies using PEG (Wilson et al., 2019), we found the dimension scores to be unidimensional in our sample (i.e., only one eigenvalue greater than one at all three time points), supporting our use of only the overall writing quality score in analyses. PEG overall scores have been found to predict statewide English language arts test scores in upper elementary grades with good to excellent classification accuracy for identifying struggling writers (Wilson, 2018).

STAAR writing test

The State of Texas Assessments of Academic Readiness (STAAR) Writing Test (Texas Education Agency, 2012a, 2012b) is a statewide assessment administered in the fourth grade and beyond. In fourth grade, students write a personal narrative and expository essay and also edit and revise the text of five given stories. The two writing samples are each scored from 1 to 4 points by two independent raters on three dimensions related to overall writing quality: structural

organization, idea development, and use of conventions. The five stories are followed by a total of 28 multiple-choice questions. Scores of 0 or 1 are attributed based on the ability of students to successfully edit and revise the written text; in particular, the students are asked to add or delete information, improve the connections between two sentences within paragraphs, and examine appropriate word selection. For the spring 2012 administration, the STAAR Writing Test composite scaled score ranged from 788 to 6517 with a mean of 3773 in 4th grade; a score of at least 3500 was the standard for satisfactory performance. The STAAR writing test demonstrated good internal consistency ($\alpha = 0.85$), interrater agreement of 61% and 98% when compositions were scored by two and three raters, respectively, and strong correlations ($r = 0.74$) with the STAAR reading test (Texas Education Agency, 2012c).

Procedures

Upon the approval of the Institutional Review Board (IRB) of the university and the participating district, the research study was presented to interested schools. Teachers administered the WE-CBM task to their students as a part of the class activities. Students completed one writing sample in September, January, and May. Students were asked to write a story in response to a story starter from the AIMSweb system (www.aimsweb.com). Students were given 1 min to think and plan the story and 3 min to complete the task. After the task was completed, teachers de-identified and provided the writing samples to the research team. Graduate students in school psychology hand-scored writing samples for WE-CBM metrics. For the purpose of the current study, graduate research assistants transcribed the writing samples from paper to electronic format. All the transcriptions were double-checked for accuracy; the project coordinator settled disagreements due to discrepancies in interpretation or difficulty transcribing due to handwriting legibility by reviewing the original writing samples. The transcribed writing samples were processed by the Coh-Metrix, ReaderBench, and GAMET desktop applications; the Coh-Metrix and ReaderBench output files were processed by the writeAlizer R package to generate quality scores. PEG scores were generated by Measurement Incorporated, the corporation maintaining PEG.

Toward the end of the same school year, as part of the state testing program, students completed the STAAR writing test. The test was administered in two 4-h sessions on consecutive days; each session required the students to write an essay and to answer questions about editing and revising given stories.

Data analyses

To address research questions regarding the criterion-related validity of automated and hand scoring, we correlated WE-CBM scores, writeAlizer quality scores (based on ReaderBench and Coh-Metrix scores; subsequently labeled as writeAlizer:RB and writeAlizer:CM), and PEG Total scores with STAAR Writing scaled scores. To determine the criterion-related validity of writeAlizer:RB and writeAlizer:CM quality scores

in combination with automated GAMET spelling and grammar scores, we entered three scores (quality, spelling, and grammar) as predictors of STAAR scaled scores in a multiple regression, with $\sqrt{R^2}$ calculated as a validity coefficient comparable to the other reported values. Consistent with a commonly-cited standard for minimally acceptable validity in WE-CBM research (McMaster & Campbell, 2008), we considered r values of ≥ 0.50 to be sufficient while also interpreting validity coefficients in relation to a more stringent standard for academic screening instruments, i.e., lower bound of the r confidence interval ≥ 0.60 (National Center on Intensive Intervention, 2018).

To determine diagnostic accuracy when using these scores to predict the attainment of satisfactory performance on the STAAR writing test, we conducted receiver operating characteristic (ROC) curve analyses, with area under the curve (AUC) statistics serving as an overall indicator of diagnostic accuracy. To determine diagnostic accuracy for writeAlizer:RB and CM quality scores combined with GAMET grammar and spelling scores, we first entered the three predictors (quality, spelling, and grammar) in a logistic regression predicting satisfactory performance on STAAR, and then used the predicted probabilities of satisfactory performance from the logistic regression in ROC analyses. We interpreted AUC values based on guidelines recommended for screening instruments (Smolkowski et al., 2016): poor diagnostic utility: AUC < 0.75; reasonable: $0.75 <$ AUC < 0.85; very good: $0.85 <$ AUC < 0.95; and excellent: AUC > 0.95.

In addition to interpreting the magnitude of individual validity coefficients, answering the research questions required comparisons of correlation coefficients and AUC values. We used Meng et al.'s (1992) z test for differences between dependent correlations with one variable in common, as implemented in the cocor R package (Diedenhofen & Musch, 2015). To compare AUC values, we used the following formula to calculate the standard error of the difference between two AUC values (Hanley & McNeil, 1983):

$$SE(AUC_1 - AUC_2) = \sqrt{SE^2(AUC_1) + SE^2(AUC_2)} \tag{1}$$

and then used these values to conduct two-tailed asymptotic z tests.

All analyses were conducted separately by time point and based on averaged scores across the three time points in R 3.6.1 (R Core Team, 2019). All students ($n = 140$) had complete data on the STAAR writing test; 130, 133, and 134 students had complete data on metrics calculated from the fall, winter, and spring writing samples, respectively. To handle missing data under the assumption that data are missing at random, analyses were based on 1000 multiply imputed datasets using the mice (van Buuren & Groothuis-Oudshoorn, 2011) and miceadds (Robitzsch & Grund, 2020) packages in R.

Results

Means and standard deviations by assessment time point for WE-CBM metrics, automated text evaluation scores, and the STAAR writing assessment are presented in Table 1.

Table 1 Means and standard deviations of scores by time point

Measure	Fall		Winter		Spring	
	M	*SD*	*M*	*SD*	*M*	*SD*
TWW	36.57	11.94	41.61	12.12	41.27	14.38
WSC	33.71	11.40	38.99	11.87	38.98	14.28
CWS	27.25	11.32	32.28	11.52	34.33	14.06
CIWS	14.20	13.67	18.58	16.18	23.20	16.17
writeAlizer:RB	0.48	237.04	13.32	208.41	19.67	256.57
writeAlizer:CM	– 17.49	258.60	– 14.60	216.97	– 3.82	247.03
Spelling	0.07	0.06	0.07	0.06	0.06	0.06
Grammar	0.01	0.02	0.01	0.01	0.01	0.01
PEG total	14.74	3.38	15.44	3.98	15.20	3.67
STAAR scaled score					3850.34	438.12
STAAR satisfactory					0.81	0.40

$n=140$. All students had complete data on the STAAR. For the other measures, 130, 133, and 134 students had complete data in the fall, winter, and spring, respectively

TWW Total Words Written; *WSC* Words Spelled Correctly; *CWS* Correct Word Sequences; *CIWS* Correct Minus Incorrect Word Sequences; *writeAlizer:RB* writeAlizer based on ReaderBench scores; *writeAlizer:CM* writeAlizer based on Coh-Metrix scores; *PEG* Project Essay Grade; *STAAR* State of Texas Assessments of Academic Readiness

Criterion-related validity

Validity coefficients (*r*s), and their 95% confidence intervals, for scores at each time point in relation to STAAR scaled scores are presented in Table 2. These coefficients demonstrate several key trends. First, nearly all the coefficients were below the recommended $r \geq 0.50$ guideline for sufficient WE-CBM validity (McMaster & Campbell, 2008). Only CIWS at the winter time point ($r=0.53$) exceeded this minimal standard.

Second, WE-CBM indicators that consider more aspects of language (CWS and CIWS) had higher validity coefficients than simpler indicators that only consider word writing productivity out of context (TWW and WSC). At the fall time point, the validity coefficient for CWS ($r=0.48$) was higher than the coefficients for TWW ($r=0.31$; $z=3.90$, $p<0.001$), and WSC ($r=0.33$; $z=3.94$, $p<0.001$). Similarly, the fall coefficient for CIWS ($r=0.49$) was higher than the coefficients for TWW ($r=0.31$; $z=2.33$, $p=0.020$), and WSC ($r=0.33$; $z=2.18$, $p=0.029$). These same contrasts were all statistically significant at the winter time point ($p<0.001$). At spring, CWS ($r=0.43$) had a higher validity coefficient than TWW ($r=0.33$; $z=2.65$, $p=0.008$) and WSC ($r=0.35$; $z=2.65$, $p=0.008$), but the CIWS coefficient ($r=0.42$), although higher, was not statistically different than the TWW ($z=1.30$, $p=0.194$) or WSC ($z=1.20$, $p=0.229$) coefficients.

Third, validity coefficients for the best-performing WE-CBM indicator (CIWS), automated writeAlizer:RB and CM scores, and automated PEG scores were comparable. At fall, although CIWS scores ($r=0.49$) had the highest

Table 2 Validity coefficients for scores at a single time point

Measure	Time point	STAAR scaled score			STAAR satisfactory		
		r	95% CI for *r*		AUC	95% CI for AUC	
			LL	*UL*		*LL*	*UL*
TWW	Fall	.31	.14	.46	.66	.55	.77
	Winter	.11 *ns*	−.06	.28	.52	.39	.64
	Spring	.33	.17	.48	.76	.67	.85
WSC	Fall	.33	.17	.48	.68	.58	.79
	Winter	.17	.01	.34	.56	.44	.69
	Spring	.35	.19	.49	.77	.69	.86
CWS	Fall	.48	.33	.60	.79	.70	.87
	Winter	.46	.31	.59	.76	.67	.85
	Spring	.43	.28	.56	.81	.73	.89
CIWS	Fall	.49	.33	.60	.79	.71	.88
	Winter	.53	.39	.64	.86	.79	.93
	Spring	.42	.27	.55	.82	.74	.89
wA:RB	Fall	.45	.31	.58	.76	.67	.85
	Winter	.42	.26	.55	.69	.58	.80
	Spring	.47	.33	.59	.82	.74	.89
wA:CM	Fall	.45	.30	.58	.74	.64	.83
	Winter	.49	.35	.61	.71	.61	.81
	Spring	.47	.32	.59	.84	.76	.91
PEG Total	Fall	.37	.21	.51	.72	.62	.82
	Winter	.45	.30	.57	.83	.75	.91
	Spring	.36	.20	.50	.76	.67	.85
wA:RB + Sp. + Gr.[a]	Fall	.48	.34	.60	.76	.67	.85
	Winter	.48	.34	.60	.79	.71	.88
	Spring	.48	.33	.60	.82	.75	.90
wA:CM + Sp. + Gr.[a]	Fall	.48	.34	.61	.76	.67	.85
	Winter	.54	.40	.65	.80	.72	.89
	Spring	.47	.33	.60	.85	.78	.92

All correlations are statistically significant at $p < .05$ except where indicated by *ns*. $N = 140$

AUC area under the receiver operating characteristic curve. *STAAR* State of Texas Assessments of Academic Readiness; *TWW* Total Words Written; *WSC* Words Spelled Correctly; *CWS* Correct Word Sequences; *CIWS* Correct Minus Incorrect Word Sequences; *wA:RB* writeAlizer with ReaderBench; *wA:CM* writeAlizer with Coh-Metrix; *PEG* Project Essay Grade; *Sp.* Spelling; *Gr.* Grammar

[a]Estimates are based on multiple regression ($r = \sqrt{R^2}$), see Table 3 for more details

validity coefficient, with writeAlizer:RB and CM scores at $r = 0.45$ for both and PEG total scores at $r = 0.37$, the correlations were not significantly different ($z = 0.57$, $p = 0.57$; $z = 0.64$, $p = 0.52$; and $z = 1.75$, $p = 0.08$, respectively). CIWS coefficients were also not statistically different from automated coefficients at the winter and spring time points.

Diagnostic accuracy

AUC values and their 95% confidence intervals for scores at each time point in relation to achieving satisfactory performance on the STAAR test are presented in Table 2. The AUC values demonstrate key trends similar to the criterion-related validity results. First, nearly all AUC values were below the 0.85 threshold suggested as very good diagnostic accuracy for a screening instrument (Smolkowski et al., 2016). Only CIWS at the winter time point met that standard (AUC = 0.86).

Second, there was a general trend of higher AUC values for the more complex WE-CBM scores (CWS and CIWS) in comparison to simpler WE-CBM scores (TWW and WSC), although the differences were not statistically significant at all time points. At the winter time point, AUC values for TWW (0.66) and WSC (0.68) were significantly lower than AUC values for CWS (0.76; $z = -3.95$, $p < 0.001$, and $z = -3.09$, $p = 0.002$, respectively) and CIWS (0.86; $z = -6.11$, $p < 0.001$, and $z = -5.09$, $p < 0.001$, respectively). By contrast, at the spring time point, AUC values for TWW (0.76) and WSC (0.77) were not statistically different than values for CWS (0.81) and CIWS (0.82).

Third, although generally lower, AUC values for writeAlizer:RB, writeAlizer:CM, and PEG were not statistically different than AUC values for CIWS at the fall and spring time points. At winter, AUCs for CIWS (0.86) and PEG (0.83) were comparable, $z = 0.60$, $p = 0.55$; however, the AUC for CIWS was higher than for writeAlizer:RB (AUC = 0.69; $z = 3.05$, $p = 0.002$) and writeAlizer:CM (AUC = 0.71; $z = 2.77$, $p = 0.006$).

Adding spelling and grammar to writealizer

To determine if adding indicators of spelling and grammar to writeAlizer quality scores would improve criterion-related validity and diagnostic accuracy, r (based on $\sqrt{R^2}$ from a multiple regression for comparability with prior analyses) and AUC values were calculated with writeAlizer, spelling, and grammar scores all entered as predictors. As demonstrated in the regression results presented in Table 3, grammar did not meaningfully contribute to the models at any time point, as evidenced by small and non-statistically significant β values; however, spelling did meaningfully contribute to the models at the winter time point for writeAlizer:RB and CM scores. Adding the automated spelling and grammar scores resulted in modest improvements in winter validity coefficients and AUC values for writeAlizer:RB + spelling + grammar ($r = 0.48$; AUC = 0.82) compared to RB alone ($r = 0.42$; AUC = 0.69) and for writeAlizer:CM + spelling + grammar ($r = 0.54$; AUC = 0.80) compared to CM alone ($r = 0.49$; AUC = 0.71). As a result, differences in winter AUCs between CIWS and writeAlizer:RB and CM were no longer statistically significant with spelling and grammar added to the writeAlizer models, $z = 0.85$, $p = 0.40$, and $z = 1.10$, $p = 0.27$, respectively.

Table 3 Multiple regression models predicting STAAR scaled scores

Variable	Fall		Winter		Spring		All (Averaged)	
	β	*SE*	β	*SE*	β	*SE*	β	*SE*
ReaderBench models								
wA:RB	.43***	.08	.32***	.08	.45***	.08	.46***	.07
Spelling	−.15	.08	−.26**	.08	−.06	.08	−.19*	.08
Grammar	.07	.08	.02	.08	.03	.08	.03	.07
R^2	.23		.23		.23		.32	
Coh-metrix models								
wA:CM	.43***	.08	.40***	.08	.44***	.08	.48***	.08
Spelling	−.17*	.08	−.24**	.08	−.08	.08	−.20*	.08
Grammar	.08	.08	.00	.08	.02	.08	.03	.07
R^2	.23		.29		.22		.34	

$N = 140$. *STAAR* State of Texas Assessments of Academic Readiness; *wA:RB* writeAlizer based on ReaderBench scores; *wA:CM* writeAlizer based on Coh-Metrix scores

*$p < .05$. **$p < .01$. ***$p < .001$

Averaging scores across three writing samples

In general, validity coefficients were improved when based on the average of scores across the three time points compared to scores from single time points. Coefficients for CWS ($r = 0.56$), CIWS ($r = 0.59$), writeAlizer:RB ($r = 0.54$), and writeAlizer:CM ($r = 0.55$) were all above the McMaster and Campbell (2008) threshold of $r \geq 0.50$, with PEG narrowly missing the threshold at $r = 0.49$. The coefficient for CIWS was not statistically different from writeAlizer:RB ($z = 1.02$, $p = 0.31$) or writeAlizer:CM ($z = 0.72$, $p = 0.47$), but was statistically greater than the PEG coefficient ($z = 2.05$, $p = 0.040$). There were no statistically significant differences among the writeAlizer:RB, writeAlizer:CM, or PEG coefficients.

For diagnostic accuracy, CIWS (AUC = 0.89) continued to be the only WE-CBM indicator with AUC above the Smolkowski et al. (2016) threshold of 0.85 for very good screening diagnostic accuracy; however, writerAlizer:RB and CM scores combined with indicators of spelling and grammar (AUC = 0.85 and 0.86, respectively) also were above this threshold. There were no statistically significant differences among the AUC values for CIWS, writerAlizer:RB, writeAlizer:CM, writeAlizer:RB + spelling + grammar, writeAlizer:CM + spelling + grammar, or PEG (Table 4).

Table 4 Validity coefficients for scores averaged over three time points

Measure	STAAR scaled score			STAAR proficiency		
	r	95% CI for *r*		AUC	95% CI for AUC	
		LL	*UL*		*LL*	*UL*
TWW	.32	.16	.46	.69	.58	.79
WSC	.36	.21	.50	.73	.63	.83
CWS	.56	.43	.66	.84	.77	.91
CIWS	.59	.47	.69	.89	.83	.95
wA:RB	.54	.41	.65	.81	.73	.89
wA:CM	.55	.42	.66	.82	.74	.89
PEG Total	.49	.35	.61	.83	.75	.90
wA:RB + Sp. + Gr.[a]	.57	.44	.67	.85	.78	.92
wA:CM + Sp. + Gr.[a]	.58	.46	.68	.86	.79	.92

$N = 140$. All correlations are statistically significant at $p < .05$

AUC area under the receiver operating characteristic curve. *STAAR* State of Texas Assessments of Academic Readiness; *TWW* Total Words Written; *WSC* Words Spelled Correctly; *CWS* Correct Word Sequences; *CIWS* Correct Minus Incorrect Word Sequences; *wA:RB* writeAlizer with ReaderBench; *wA:CM* writeAlizer with Coh-Metrix; *Sp.* Spelling; *Gr.* Grammar; *PEG* Project Essay Grade

[a]Estimates are based on multiple regression ($r = \sqrt{R^2}$), see Table 3 for more details

Discussion

The primary purpose of this study was to compare the criterion-related validity and diagnostic accuracy of commercial and free automated text evaluation tools, relative to hand-scored WE-CBM, for universal screening of writing skills in fourth grade students. Overall, criterion-related validity and diagnostic accuracy were comparable for the commercial tool (PEG), the free tool (writeAlizer using ReaderBench or Coh-Metrix scores), and the best performing WE-CBM score (CIWS). Consistent with the findings of a prior meta-analysis (Romig et al., 2017), we found validity to be stronger for more complex WE-CBM scores (CIWS and CWS) than simpler scores (TWW and WSC). At the winter time point, diagnostic accuracy was greater for CIWS and PEG, relative to writeAlizer, but differences in diagnostic accuracy were no longer statistically significant when writeAlizer scoring was expanded to include automated spelling and grammar scores. When predicting state writing test scores, automated spelling scores, but not automated grammar scores, were statistically significant predictors in combination with writeAlizer scores based on ReaderBench or CohMetrix. These findings suggest that there may be limited value in including the automated grammar scores in the writeAlizer model; however, replication in other samples and with other writing criterion measures is needed to determine optimal weightings for automated spelling and grammar scores before fully integrating them in a revised writeAlizer scoring model.

In general, when evaluating scores based on 3-min screening samples at one time point, other than CIWS at the winter time point, criterion-related validity and diagnostic accuracy were below recommended performance thresholds for screening instruments ($r \geq 0.50$ considered sufficient in McMaster & Campbell, 2008; AUC > 0.85 considered very good in Smolkowski et al., 2016). By averaging scores across three screening samples, validity coefficients and diagnostic accuracy were improved, with performance for more complex WE-CBM scores and automated tools approaching or meeting the above recommended performance thresholds. Although it is not typical to average scores across three screening time points, we used this analytic approach to approximate the use of multiple samples per student and time point given that only one sample per time point was collected in the current study.

The current findings extend prior research in three key areas. First, the findings of greater validity and diagnostic accuracy for all scoring approaches when based on three versus one screening sample per student build on prior generalizability theory studies demonstrating that multiple writing samples are needed for adequate reliability (Keller-Margulis et al., 2016; Kim et al., 2017). Second, the findings of comparable validity and diagnostic accuracy across complex WE-CBM scoring, PEG scoring (commercial), and writeAlizer with ReaderBench (open source) and Coh-Metrix (free) scores build on prior work showing (a) only modest improvements in diagnostic accuracy on a state writing assessment for commercial text evaluation relative to simple WE-CBM scoring of 30-min samples (Wilson et al., 2016) and (b) comparable relations to holistic quality for 7-min samples scored with free automated text evaluation and complex WE-CBM metrics (Mercer et al., 2019). These findings illustrate that it may be difficult to improve validity of automated text evaluation beyond complex WE-CBM scoring; instead, investigating the optimal number of writing samples and writing duration for optimal reliability may yield more benefits in terms of improved validity. Last, although not directly examined in the current study, the findings suggest potential feasibility benefits to using automated text evaluation for screening. The time investment when using automated approaches is in the transcription of samples, as opposed to scoring when using traditional approaches, and no transcription would be necessary if students compose samples electronically. Our findings of comparable performance for free compared to commercial automated text evaluation may further improve feasibility by reducing the cost of scoring software.

In general, the results of this study suggest promise for the use of automated text evaluation approaches for universal screening in terms of validity and feasibility. Despite these promising findings, there is still considerable room for improvement. None of the validity coefficients would meet the National Center for Intensive Intervention's (2018) "convincing evidence" validity standard for academic screening instruments of the lower bound of the confidence interval ≥ 0.60, but for averages across three samples, diagnostic accuracy values for complex WE-CBM scores and automated scores met the "partially convincing evidence" standard of the lower bound of the AUC confidence interval ≥ 0.70, and CIWS would meet the "convincing evidence" standard of ≥ 0.80 for the lower

bound. Of note, though, currently no written expression screening tools are listed on NCII's tool chart, highlighting the need for continued work in this area.

Limitations and future directions

The results of this study should be considered in the context of several limitations. First, although consistent with recommended WE-CBM procedures (e.g., Hosp et al., 2016), the 3-min writing duration used in this study may have attenuated the validity estimates for all scoring approaches. In addition, averaging scores across multiple screening time points is not a practical approach, but it did provide some insight regarding the potential value of basing scores on multiple screening samples. Future research is needed to identify the number of samples and writing duration needed for optimal reliability and validity, both for traditional WE-CBM and automated scoring approaches. Also, we only examined writing samples generated in response to narrative prompts. Although narrative is the most frequently used genre in WE-CBM (see McMaster & Espin, 2007), other genres are also emphasized in writing instruction (Philippakos et al., 2015), and obtaining screening samples across multiple genres could potentially improve validity. Associated with the sample, there was a fairly significant percentage of the sample identified as English Learners. The district where these data were collected used a late-exit transitional bilingual model of instruction where students received primarily Spanish language instruction with some English in kindergarten. That percentage of instruction across languages shifted 50% of each language in third grade with only English as a second language support services and all academic instruction in English by fifth grade. All students in this sample were instructed in English and tested in English, however it is not possible to rule out the potential impact of language status on performance. Last, we focused on the use of automated scoring for screening to identify students at risk for poor performance, but future work should also investigate automated scoring for progress monitoring because the ability to monitor progress in response to changes in instruction or intervention is an essential use of CBM.

Conclusion

There is a practical need for tools that can be efficiently used to identify students at risk for poor performance in written expression. The current findings indicate that automated scoring approaches to scoring writing are a promising alternative to more complex hand-scored WE-CBM metrics, with potential benefits in scoring efficiency and feasibility. Given that validity and diagnostic accuracy were comparable across complex WE-CBM scoring and both free and commercial automated text evaluation programs, the optimal number of writing samples used for screening needs more attention in future research.

Acknowledgements This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A190100 awarded to the University of Houston (PI – Milena

Keller-Margulis). The opinions expressed are those of the authors and do not represent views of the Institute or thae U.S. Department of Education.

References

Allen, L., Dascalu, M., McNamara, D. S., Crossley, S., & Trausan-Matu, S. (2016). Maodeling individual differences among writers using ReaderBench. In L. Gómez Chova, A. López Martínez, & I. Candel Torres (Eds.), *EDULEARN16 proceedings.* (pp. 5269–5279). IATED Academy.

Botarleanu, R. M., Dascalu, M., Sirbu, M. D., Crossley, S. A., & Trausan-Matu, S. (2019). ReadMEGenerating personalized feedback for essay writing using the ReaderBench framework. In H. Knoche, E. Popescu, & A. Cartelli (Eds.), *Smart learning ecosystems and regional development 2018.* (pp. 133–145). Springer.

Cook, B. G., Lloyd, J. W., Mellor, D., Nosek, B. A., & Therrien, W. J. (2018). Promoting open science to increase the trustworthiness of evidence in special education. *Exceptional Children, 85*, 104–118. https://doi.org/10.1177/0014402918793138.

Crossley, S. A., Bradfield, F., & Bustamante, A. (2019). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research, 11*, 251–270. https://doi.org/10.17239/jowr-2019.11.02.01.

Dascalu, M., Crossley, S. A., McNamara, D. S., Dessus, P., & Trausan-Matu, S. (2018). Please ReaderBench this text: A multi-dimensional textual complexity assessment framework. In S. D. Craig (Ed.), *Tutoring and intelligent tutoring systems.* (pp. 251–271). Nova Science.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219–232. https://doi.org/10.1177/001440298505200303.

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE, 10*, 1–12. https://doi.org/10.1371/journal.pone.0121945.

Espin, C., Shin, J., Deno, S. L., Skare, S., Robinson, S., & Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *The Journal of Special Education, 34*, 140–153. https://doi.org/10.1177/002246690003400303.

Espin, C. A., Scierka, B. J., Skare, S., & Halverson, N. (1999). Criterion-related validity of curriculum-based measures in writing for secondary school students. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 15*, 5–27. https://doi.org/10.1080/105735699278279.

Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Naquin, G. M., & Slider, N. J. (2002). Moving beyond total words written: The reliability, criterion validity, and time cost of alternate measures for curriculum-based measurement in writing. *School Psychology Review, 31*, 477–497.

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal, 115*, 210–229. https://doi.org/10.1086/678293.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148*, 839–843. https://doi.org/10.1148/radiology.148.3.6878708.

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). *The ABCs of CBM: A practical guide to curriculum-based measurement* (2nd ed.). Guilford.

Keller-Margulis, M. A., Mercer, S. H., & Thomas, E. L. (2016). Generalizability theory reliability of written expression curriculum-based measurement in universal screening. *School Psychology Quarterly, 31*, 383–392. https://doi.org/10.1037/spq0000126.

Kim, Y. G., Schatschneider, C., Wanzek, J., Gatlin, B., & Al Otaiba, S. (2017). Writing evaluation: Rater and task effects on the reliability of writing scores for children in Grades 3 and 4. *Reading and Writing: An Interdisciplinary Journal, 30*, 1287–1310. https://doi.org/10.1007/s11145-017-9724-6.

Malecki, C. K., & Jewell, J. (2003). Developmental, gender, and practical considerations in scoring curriculum-based measurement writing probes. *Psychology in the Schools, 40*, 379–390. https://doi.org/10.1002/pits.10096.

McMaster, K. L., & Campbell, H. (2008). New and existing curriculum-based writing measures: Technical features within and across grades. *School Psychology Review, 37*, 550–556.

McMaster, K. L., & Espin, C. A. (2007). Technical features of curriculum-based measurement in writing. *The Journal of Special Education, 41*, 68–84. https://doi.org/10.1177/00224669070410020301.

McMaster, K. L., Lembke, E. S., Shin, J., Poch, A. L., Smith, R. A., Jung, P., Allen, A. A., & Wagner, K. (2020). Supporting teachers' use of data-based instruction to improve students' early writing skills. *Journal of Educational Psychology, 112*, 1–21. https://doi.org/10.1037/edu0000358.

Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*, 172–175. https://doi.org/10.1037/0033-2909.111.1.172.

Mercer, S. H. (2020). *writeAlizer: Generate predicted writing quality and written expression CBM scores*. (Version 1.2.0) [Computer software]. https://github.com/shmercer/writeAlizer/.

Mercer, S. H., Keller-Margulis, M. A., Faith, E. L., Reid, E. K., & Ochs, S. (2019). The potential for automated text evaluation to improve the technical adequacy of written expression curriculum-based measurement. *Learning Disability Quarterly, 42*, 117–128. https://doi.org/10.1177/0731948718803296.

National Center for Educational Statistics. (2012). *The nation's report card: Writing 2011*. Institute of Education Sciences, U.S. Department of Education. http://nationsreportcard.gov.

National Center on Intensive Intervention. (2018). *Academic screening tools chart rating rubric.* https://intensiveintervention.org/sites/default/files/NCII_AcademicScreening_RatingRubric_July2018.pdf.

Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective.* (pp. 43–54). Lawrence Erlbaum Associates.

Payan, A. M., Keller-Margulis, M., Burridge, A. B., McQuillin, S. D., & Hassett, K. S. (2019). Assessing teacher usability of written expression curriculum-based measurement. *Assessment for Effective Intervention, 45*, 51–64. https://doi.org/10.1177/1534508418781007.

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing, 21*, 104–111. https://doi.org/10.1016/j.asw.2014.05.001.

Perin, D. (2020). Reading, writing, and self-efficacy of low-skilled postsecondary students. In D. Perin (Ed.), *The Wiley handbook of adult literacy.* (pp. 237–260). Blackwell: Wiley.

Philippakos, Z. A., MacArthur, C. A., & Coker, D. L. (2015). *Developing strategic writers through genre instruction: Resources for grades 3–5*. Guilford.

R Core Team. (2019). *R: A language and environment for statistical computing*. (Version 3.6.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/.

Rainie, L., & Anderson, J. (2017). *Code-dependent: pros and cons of the algorithm age*. Pew Research Center. http://www.pewinternet.org/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age.

Ritchey, K. D., McMaster, K. L., Al Otaiba, S., Puranik, C. S., Kim, Y. G., Parker, D. C., & Ortiz, M. (2016). Indicators of fluent writing in beginning writers. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and applications.* (pp. 21–66). Springer.

Robitzsch, A., & Grund, S. (2020). *miceadds: Some additional multiple imputation functions, especially for 'mice'*. (Version 3.9–14) [Computer software]. https://CRAN.R-project.org/package=miceadds.

Roebuck, D. B., Sightler, K. W., & Brush, C. C. (1995). Organizational size, company type, and position effects on the perceived importance of oral and written communication skills. *Journal of Managerial Issues, 7*, 99–115.

Romig, J. E., Therrien, W. J., & Lloyd, J. W. (2017). Meta-analysis of criterion validity for curriculum-based measurement in written language. *The Journal of Special Education, 51*, 72–82. https://doi.org/10.1177/0022466916670637.

Smolkowski, K., Cummings, K. D., & Strycker, L. (2016). An introduction to the statistical evaluation of fluency measures with signal detection theory. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and applications.* (pp. 187–221). Springer.

Stevens, B. (2005). What communication skills do employers want? Silicon Valley recruiters respond. *Journal of Employment Counseling, 42*, 2–9. https://doi.org/10.1002/j.2161-1920.2005.tb00893.x.

Texas Education Agency. (2012a). *State of Texas assessments of academic readiness: Grade 4 expository scoring guide spring 2012.* https://tea.texas.gov/sites/default/files/staar-g4-ExpScorGde-spr2012.pdf.

Texas Education Agency. (2012b). *State of Texas assessments of academic readiness: Grade 4 personal narrative scoring guide spring 2012.* https://tea.texas.gov/sites/default/files/staar-g4Wtg-PerNarrScoreGde-Spr2012.pdf.

Texas Education Agency. (2012c). *Technical digest 2011–2012.* https://tea.texas.gov/student-assessment/testing/student-assessment-overview/technical-digest-2011-2012.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*, 1–67. https://doi.org/10.18637/jss.v045.i03.

Wilson, J. (2018). Universal screening with automated essay scoring: Evaluating classification accuracy in grades 3 and 4. *Journal of School Psychology, 68*, 19–37. https://doi.org/10.1016/j.jsp.2017.12.005.

Wilson, J., Chen, D., Sandbank, M. P., & Hebert, M. (2019). Generalizability of automated scores of writing quality in Grades 3–5. *Journal of Educational Psychology, 111*, 619–640. https://doi.org/10.1037/edu0000311.

Wilson, J., Olinghouse, N. G., McCoach, D. B., Santangelo, T., & Andrada, G. N. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing, 27*, 11–23. https://doi.org/10.1016/j.asw.2015.06.003.

Wilson, J., Roscoe, R., & Ahmed, Y. (2017). Automated formative writing assessment using a levels of language framework. *Assessing Writing, 34*, 16–36. https://doi.org/10.1016/j.asw.2017.08.002.